

DOCUMENT RESUME

ED 220 527

TM 820 549

AUTHOR Baldwin, Lee; Medley, Donald M.
TITLE A Comparison of Analysis of Covariance to Within-Class Regression in the Analysis of Non-Equivalent Groups.
PUB DATE Mar 82
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Analysis of Covariance; Hypothesis Testing; *Matched Groups; *Quasiexperimental Design; Reading Achievement; *Regression (Statistics); Research Design; Statistical Analysis; *Statistical Significance; Statistical Studies
IDENTIFIERS Monte Carlo Methods; *Within Class Regression

ABSTRACT

The purpose of this study was to compare the results of an analysis of covariance design to a method developed by the authors called within-class regression. Application of this statistical method can be to any quasiexperimental design with a large number of groups. A test of statistical significance to use with the within-class regression procedure was also developed using an analysis of variance model to estimate between- and within-groups components of variance. The two methods were compared using Monte Carlo data and reading achievement data. Achievement Development Scale scores from the California Achievement Tests were collected in grades two through six in May of 1978 and 1979. The results of the analysis under various conditions indicated that within-class regression is a less biased method of analysis, and that different conclusions concerning treatment effects may be reached depending on the method of analysis used. (Author/PN)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED220527

A Comparison of Analysis of Covariance
to Within-Class Regression in the Analysis
of Non-Equivalent Groups

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.
☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

Lee Baldwin

P. K. Yonge Laboratory School
University of Florida
Gainesville, Florida 32611

Donald M. Medley

Department of Educational
Research and Evaluation
University of Virginia
Charlottesville, Virginia 22903

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

L. Baldwin

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

A Paper Presented at the Annual Convention
of the
American Educational Research Association
New York, New York

March 19, 1982

TM 820 549

A Comparison of Analysis of Covariance
to Within-Class Regression in the Analysis
of Non-Equivalent Groups

Classical statistical designs used in educational research are often inadequate when applied to groups that have not been formed at random. In field settings, researchers must rely on statistical techniques that were developed for closely controlled laboratory experiments. Often these statistical techniques are sensitive to the violation of assumptions that occur when applied in field settings to intact groups.

The purpose of this study was to compare the results of an analysis of covariance design to a method described in this article called within-class regression. This method may be appropriate when a large number of intact groups are to be compared. Groups were not assumed to have been randomly formed and there were preexisting differences among the groups.

The methods were compared using computer-generated Monte Carlo data and data from a quasi-experimental evaluation design. Since analysis of covariance designs make rather rigorous assumptions about the nature of the data, the use of these designs often leads to biased conclusions

when applied to quasi-experimental research. If a method of data analysis can be developed in which such rigorous assumptions are not made, a less biased, more accurate analysis may be possible.

Several techniques have been suggested to compare intact groups. Raw gain scores or change scores have the advantage of being attractive on an intuitive level. The gain score is found by subtracting a pre-treatment measure from a post-treatment measure. However, gain scores have been shown to be less reliable than either of the original scores (Lord, 1963). Because of unreliability of the gain score, use of a gain score as a criterion score for change is generally not recommended.

A second method for measuring change is the use of ex post facto matching techniques. Matching techniques require that a subject from one group be matched with a subject from another group on the basis of some important concomitant variable or variables. This technique presents two problems: one logistical and one methodological. If there is no subject in the second group that can be matched with a subject in the first group, the subject from the first group must be deleted from the analysis. As the matching criteria or the number of groups become larger, the number of deleted subjects increases. This loss of subjects reduces the number of degrees of freedom for the statistical test, thus increasing the probability of a type two error. When a

large number of groups are compared, this method becomes completely impractical since the overlapping parts of the distributions on the matching variable become so small that virtually no subjects remain.

In addition to the problem of finding matches, a methodological problem results because of a regression phenomenon. If the mean scores of two groups on the matching variable are different, the upper part of the distribution of one group will overlap with only the lower part of the distribution of the other group. Because of unreliability of the measure (Campbell & Erlebacher, 1971) and other factors (Furby, 1973; Lord, 1963), subjects' scores will tend to regress toward their respective population means. This phenomenon will result in a bias favoring the group initially higher on the matching variable.

The third method for equating intact groups that are initially different is the analysis of covariance. This method became the most popular method during the 1960's and 1970's (Campbell & Erlebacher, 1971; Cronbach, Rogosa, Floden, & Price, 1976) and is currently the most used method of analyzing intact groups. However, analysis of covariance (Ancova) has also been shown to result in biased adjustments which also usually favor the group initially higher on the concomitant variable (Campbell & Boruch, 1975; Campbell & Erlebacher, 1971). Elashoff (1969) discussed the assumptions underlying Ancova and concluded that the analysis of covariance is a delicate instrument and must be used with caution or biases may become confounded with treatment effects.

Analysis of covariance includes a linear regression of the posttest on the pretest to partial out variance in the posttest that can be predicted from the pretest. The adjusted scores are then compared among groups. When analysis of covariance is used, it is assumed that the regression of the posttest on the pretest is homogeneous for all groups (Burstein, Linn, & Capell, 1978; Elashoff, 1969). When the assumption of homogeneity of regression is violated, the main effect may be uninterpretable because different treatments may be more or less effective at different levels. A second assumption of analysis of covariance is that the covariate can be measured with perfect reliability (Campbell & Erlebacher, 1971; Lord, 1960). If the covariate is not perfectly reliably measured, a bias will be introduced when equating the groups. This bias will usually favor the initially higher group (Cronbach et al., 1976). Some techniques have been suggested in the literature to correct covariance adjustments for unreliability (Lord, 1960, 1969; Porter, 1971). These techniques, however, assume reliability to be equal for all groups (Campbell & Boruch, 1975). When reliability is not equal across groups, adjustments will be incorrect (Campbell & Boruch, 1975). Although some authors consider Ancova to be inadequate (Campbell & Boruch, 1975), others consider the method to be the preferred method of analysis when comparing intact groups (Cronbach et al., 1976).

Description of the Method

Within-class regression is a method of measuring a large number of intact groups such as classes. When this method is used, as with the analysis of covariance, it is assumed that a linear model represents the process for the generation of data. Given that equivalent forms of a posttest and pretest are available for each group, the common pretest-posttest quasi-experimental design is evident as diagrammed below:

X_1	O_1	X_2
X_1	O_2	X_2
X_1	O_3	X_2
.	.	.
.	.	.
.	.	.
X_1	O_k	X_2

where X_1 is the pretest for all of the groups, X_2 the posttest, and O_1 through O_k the treatments for groups 1 through k , respectively.

The data analysis, using within-class regression, involves estimating a correlation between pre- and posttest scores

within each class. Regression lines, regressing the post-test on the pretest, are then derived. This regression procedure differs from analysis of covariance in that regression lines are computed within each group instead of computing a pooled within-groups regression line which is, in effect, a composite of regression lines within groups. These individual regression lines allow an important assumption of analysis of covariance to be avoided; namely, the assumption of homogeneity of regression.

The models for the two techniques can be shown as follows. The model for analysis of covariance is:

$$Y_{ij} = \mu + \alpha_j + \beta (X_{ij} - \bar{X}_{..}) + e_{ij} \quad (1)$$

$\{i = 1, 2, \dots, n; j = 1, \dots, k\}$, where

Y_{ij} = the score on the criterion variable made by individual i in treatment group j .

μ = the grand mean on the criterion variable.

α_j = the effect on the mean due to treatment j .

$\beta(X_{ij} - \bar{X}_{..})$ = the portion of individual variability due to linear regression of Y or X .

e_{ij} = random or unexplained error.

β = the pooled within-group regression coefficient for all k groups of the criterion variable on the concomitant variable.

The model for within-class regression is the simple regression model within any one group.

$$Y_{1j} = \alpha_j + \beta_j (X_{1j} - \bar{X}_{.j}) + e_{1j} \quad \text{where (2)}$$

β_j = regression coefficient for group j .

$\bar{X}_{.j}$ = the mean of the concomitant variable within treatment group j .

The terms Y_{1j} , α_j , X_{1j} , and e_{1j} have the same meaning in equation (2) as in equation (1) representing the postscore, Y intercept, and pretest score, respectively. As can be seen from equation (2), the within-class regression model allows for a heterogeneity of regression of the criterion variable on the concomitant variable across classes.

A score Y'_z can be predicted for any pretest score of X by using the regression equation. If homoscedasticity of regression is assumed, the predicted score within a group will be the mean score for the population of scores with a pretest score of X . Further, this $\bar{Y}'_z$ score is an indication of the mean posttest score for members of a group who have a prescore of X .

Through the use of this method, regression is allowed to vary within groups and regression effects will be toward the group mean and not assumed to be toward an overall grand mean. If there is, in fact, a heterogeneity of regression, this phenomenon is preserved and not suppressed. However, the $\bar{Y}'_z$ value is not a satisfactory measure of how effective the treatment is. If a subject scores at the grand

mean of the pretest and that score is below average in one group, the score will tend to regress upward toward the group mean. A subject with the same score in another group will tend to regress down to the mean if the score is above average in that group. At this point, because the groups are not comparable, there is a regression artifact favoring the class initially higher on the pretest.

By correlating the $\bar{Y}'_z$ values at a certain level of the pretest with the mean pretest values for that group, a second regression line, showing the values of $\bar{Y}'_z$ that would be expected if all treatment effects were equal, can be derived. The second regression between means at one point on the distribution controls for differences among group means. If all treatments were equally effective at that point on the distribution, then the points would all lie on the regression line except for random fluctuations. If one treatment were more effective than average, that residual mean would lie above the regression line. Conversely, if the treatment were less effective than average, the residual mean would lie below the regression line. The residual means thus indicate the relative strength of the treatment effects.

This method of measuring treatment effects is appealing from several perspectives. As already shown, the assumption of a homogeneity of regression is not made. The different regression lines do not suppress what may be an important phenomenon, namely a treatment-level interaction. The

possibility of a treatment-by-level interaction is not merely a theoretical concern but also a practical concern in educational research.

Burstein, Linn, and Capell (1978) conceptualized the differences of regression lines that could be expected from classrooms. They contended that, in a class in which individualized instruction is emphasized, higher ability students would be expected to make more appropriate decisions and learn at a faster rate than lower ability students. In this situation, individualization could easily strengthen the relationship between entering ability and outcome. On the other hand, a teacher in a class where mastery learning is emphasized might stress that all students learn the curriculum content. Teachers in this situation might compensate for preexisting differences and direct resources, materials, and time toward those students with the poorest entering performance. Such a practice would most likely lead to flatter within-class regression lines. Using this conceptualization the individualized instructional approach would probably lead to a steeper regression line than the compensatory or mastery approach.

Analysis of covariance designs allow for only a comparison of an overall treatment effect. With the within-class regression method, the treatment effect can be analyzed at any level of the pretest. Comparison can be made at the grand mean of the pretest but can just as easily be made at a level say of one standard deviation above the mean or one standard deviation below the mean.

It also seems that the assumption of perfect reliability of the pretest may be relaxed. In analysis of covariance, groups are equated on a concomitant variable. If the variable is not measured with perfect reliability, the two groups will tend to regress to their respective population means, thus creating a regression artifact favoring, in most cases, the initially superior group. By using within-class regression, no effort is made to equate groups. Regression at the first step will be present but the regression will be toward the respective group means. At the second step, means and not individual scores are correlated, and no regression artifact should be present. These steps seem to reduce confounding that is caused when groups are equated. Since regression toward one population mean is not built into the model, it seems that a regression artifact caused by unreliability will not appear or at least will be reduced.

A Test of Significance

A test of statistical significance to use with the within-class regression procedure was also developed. A significance test must test the significance of the differences of residual mean scores at some point on the distribution of X scores. The Johnson-Neyman technique, later expanded by Pothoff, was developed to make such a test in the case of non-parallel regression lines in an experimental setting (Pothoff, 1964). This technique is closely related to an analysis of covariance procedure at a particular point on the X

distribution. The following procedure borrows somewhat from the Johnson-Neyman technique, especially in the estimate of the within-group variance.

The significance test uses an analysis of variance model to estimate between- and within-groups components of variance. The residual mean scores remaining after correlating the estimated scores at a point on the pretest distribution with the mean of the pretest scores becomes the source of estimating the between-groups sum of squares.

The formula for estimating the between-groups variance for a score (Y) when deviation scores are used is usually given as:

$$\sum_{i=1}^k \frac{(\Sigma y)^2}{n}$$

Since

$$\bar{y} = \frac{\Sigma y}{n}$$

and

$$\bar{y}n = \Sigma y,$$

the between-groups sum of squares can be estimated by substituting $\bar{y}n$ for Σy to yield the formula:

$$\sum_{i=1}^k \frac{(\bar{y}n)^2}{n}$$

which can be reduced to

$$\sum_{i=1}^k (\bar{y})^2 n.$$

By summing these squared residual sums from $i = 1$ to k , an estimate of the between-groups sum of squares is obtained. Since the residual scores are already deviation scores, no correction term is needed.

To estimate the within-groups variance, the concept of the standard error of estimate is employed. The standard error of estimate for any distribution is:

$$SE_{y|x} = S_y (1-r_{xy}^2)^{\frac{1}{2}}$$

By squaring both sides, we get a standard variance error estimate:

$$S_{y|x}^2 = S_y^2 (1-r_{xy}^2)$$

Sums of squares estimates can be derived by multiplying both sides by n , which yields:

$$SS_{y|x}^2 = SS_y^2 (1-r_{xy}^2)$$

To arrive at the total within-groups sum of squares, the sum of squares within each group is summed for a total within-groups sum of squares.

The degrees of freedom are $k - 1$ for the between-groups sum of squares and $k(n - 2)$ for the within-groups sum of squares. Calculation of mean squares and the F ratio follow the normal analysis of variance computations.

Research Design

The design of this study may be divided into two parts. In the first part, Monte Carlo data were computer-generated under varying assumptions to measure the bias of the two methods. In the second part of the study, real data were analyzed to measure the effectiveness of teachers in teaching reading. The statistical properties of the methods were observed under carefully controlled conditions in the first part of the study. In the second part of the study, the results of the methods of analysis were compared in a real quasi-experimental situation.

Part I: Monte Carlo Analysis

Random data were computer-generated under three conditions. In the first condition, the reliability of the pre- and posttests was less than perfect but equal in all classes. This condition allowed random error to enter in the data generation model. In the second condition the reliability of the measures between groups was allowed to vary. Reliability was operationalized as stability and was measured by a pretest-posttest correlation in the absence of a treatment effect. Often it is assumed that reliability is the same across groups, even in formulae that correct for unreliability (Campbell & Boruch, 1975). However, when intact groups are analyzed, this assumption becomes somewhat tenuous. In the third condition, posttest variances were allowed to vary across groups. This condition violates Ancova assumptions of homogeneity of variance.

A factor was built into the model to make the groups unequal at the outset and thus potentially bias the results of the statistical analyses. Additionally, a treatment effect was built into the model. However, since the interest in this study centered on bias and not on differing treatment effects, the treatment effect was controlled by setting it equal to zero for all groups.

Under these controlled conditions, the data generation model for the pretest and the posttest can be shown as:

$$X_{ij} = G_{.j} + \beta_{j1}T_{ij} + \beta_{j2}E_{ij} \quad (3)$$

$$Y_{ij} = \alpha_{.j} + G_{.j} + \beta_{j1}T_{ij} + \beta_{j2}E'_{ij}, \quad \text{where} \quad (4)$$

X_{ij} = the pretest score for person i in treatment j .

Y_{ij} = the posttest score for person i in group j .

$\alpha_{.j}$ = the treatment effect across subjects in group j .

$G_{.j}$ = a score common to all members of group j .

G is that portion of true score variance producing group differences.

T_{ij} = the common factor score between pretest and posttest for subject ij . It represents that component of true score variance accounting for individual differences in the prescore between groups. In this simulated model,

neither pretest nor posttest contain true score components not shared with the other.

E_{ij} & E'_{ij} = separate error scores on the pretest and posttest for subject ij . They represent random measurement error. E_{ij} and E'_{ij} are different and cause the pretest and posttest observed scores for a subject to be different.

β_{j1} & β_{j2} = weighting coefficients of the true score and error components for group j . The ratio of the true score weighting to the total weighting in the pretest and posttest determined the stability of the measures. If

$$\frac{\beta_{j1}}{\beta_{j1} + \beta_{j2}} = 1, \quad (5)$$

then β_{j2} must equal 0. This would mean that the test is only composed of a true score component and reliability would be perfect.

This model for data generation allowed for a treatment effect, a factor causing preexisting differences, a true score component, and separate error scores in the pretest and the posttest. Also the reliability between groups was allowed to vary for the second condition. In order to simplify the interpretation of the results, the treatment effect (α_j) was set equal to zero for all treatment groups, $j = 1, 2, \dots, k$. Because the treatment effect was zero, the term was deleted from equation (4). As a result, any treatment effects indicated by

the methods of analysis were pseudo-effects and were considered regression artifacts.

The terms T_{ij} , E_{ij} , and E'_{ij} were generated in a normal distribution and all three terms were standardized scores with 0 means and unit variances. Setting the means and standard deviations to be equal did not change the results of the analysis but facilitated interpretation and restricted the model to two variables: preexisting differences between groups and differences in reliability among groups. This model is the same model used by Campbell and Erlebacher (1970) except for the addition of the weighting coefficients β_{j1} and β_{j2} .

The third condition required the addition of a treatment-by-level interaction term to the model. In this condition, reliability was held constant across groups, but preexisting differences remained. The effect of a violation of the assumption of analysis of covariance that there is no treatment-by-level interaction was tested.

If the overall treatment effect (α_j) is still held equal to 0 for all groups, then, under the condition of the existence of a treatment-level interaction, the standard deviation of the posttest distribution would change relative to the pretest distribution. In a classroom setting, if a teacher were more effective with lower achieving students and less effective with higher achieving students, the lower achieving students would tend to regress upward toward the mean of the class. The above-average students would, however,

tend to regress downward toward the mean of the class. If this phenomenon occurred, the standard deviation of the posttest would be smaller than the standard deviation of the pretest.

In the event of a teacher effect where the more able students learn at a faster rate than the less able students, again under the constraint of no overall treatment effect, the more able students would tend to move upward away from the mean and the less able students would tend to move downward away from the mean. In this case the posttest standard deviation would become greater than the pretest standard deviation. These results are under the strict constraint of no overall treatment effects. In a real setting, some treatment effects would probably be present and the low achieving students would not actually regress downward away from the mean of their class but only make relatively less progress. The standard deviation would, however, still increase at the posttest.

Since the regression coefficient is equal to the correlation between X (pretest) and Y (posttest), multiplied by the ratio of the standard deviation of Y over the standard deviation of X, the slope will be greater if $\frac{S_y}{S_x} > 1$ than if $\frac{S_y}{S_x} < 1$. Thus, the teacher's class whose less able pupils learn relatively more should have a flatter regression line than the teacher whose more able students learn more when all conditions except the standard deviation of the posttest are held constant.

The model for the generation of data under the condition of a treatment-by-level interaction was:

$$X_{1j} = G_{.j} + \beta_1 T_{1j} + \beta_2 E_{1j} + \beta_{j3}(\beta_1 T_{1j} + \beta_2 E_{1j}) \quad (6)$$

$$Y_{1j} = G_{.j} + \beta_1 T_{1j} + \beta_2 E'_{1j} + \beta'_{j3}(\beta_1 T_{1j} + \beta_2 E'_{1j}) \quad (7)$$

All terms are the same as in equations 3 and 4 except the addition of the interaction term. β_1 and β_2 are constrained so that they remain constant across groups. This constraint holds the reliability of the scores constant across groups. Since the T_{1j} , E_{1j} , and E'_{1j} scores are z scores with 0 means, any scores below the mean are negative scores; any scores above the mean are positive scores. In this case the β_{j3} factor causes below average scores to move downward away from the mean to the same extent that it moves above average scores upward away from the mean when $\beta_{j3} < \beta'_{j3}$. When $\beta_{j3} > \beta'_{j3}$, both below average and above average scores regress toward the mean on the posttest. The equations can be simplified somewhat if $\beta_{j3}(\beta_1 T_{1j} + \beta_2 E_{1j})$ is set equal to 0. Setting this term equal to zero only has the effect of setting the treatment-by-level interaction to 0 at the pretest. Making this simplification, the equations may be rewritten as:

$$X_{1j} = G_{.j} + \beta_1 T_{1j} + \beta_2 E_{1j} \quad (8)$$

$$Y_{1j} = G_{.j} + \beta_1 T_{1j} + \beta_2 E'_{1j} + \beta'_{j3}(\beta_1 T_{1j} + \beta_2 E'_{1j}) \quad (9)$$

After this simplification, β'_{j3} will still affect the standard deviation as before. Since the pretest interaction term has been set equal to zero, if β'_{j3} is greater than zero, the posttest score will become larger if $(\beta_1 T_{1j} + \beta_2 E'_{1j})$ is greater than zero. If β'_{j3} is less than zero, the absolute value of any posttest score will become smaller. If β'_{j3} is less than zero, the standard deviation will tend to shrink until $\beta'_{j3} = -1$. At this point, regression is total and all scores will be expected to regress to the mean.

After generating data under both models, the results were analyzed using both methods of statistical analysis. Since no treatment effect was built into the model, neither one of the statistical methods should indicate an overall treatment effect. Any deviation away from a no-treatment effect was a measure of a bias in the method of analysis caused by regression artifacts.

Data for all three conditions of this part of the study were generated according to the data matrix shown in Table 1. There were 20 classes with 500 subjects per class. Each subject had a pretest score and a posttest score.

Insert Table 1 about here

In the first condition, the correlation between the pre- and posttest was set equal to .75 except for random fluctuations. In the second condition, the correlation coefficient was allowed to vary around a mean correlation coefficient of .75. The 20 coefficients were generated in a normal distribution with a

standard deviation of .1. These parameters yielded correlation coefficients ranging from .64 to .93. A mean coefficient of .75 was considered a reasonable stability coefficient to expect for achievement test data.

The β'_{j3} factor in the third condition was also allowed to vary. The mean value of the coefficient β'_{j3} was set at 0 with a standard deviation of .5. These parameters caused the weighting coefficients to range from a value of -0.89 to 1.26. Twelve coefficients were positive and eight were negative. The values generated caused the standard deviations of the posttest to vary in relation to the pretest from more than doubling the standard deviation (1.26) to removing most of the variability in the posttest scores (-0.89).

Part 2: Reading Achievement Test Data Analysis

The same two methods were used to analyze reading achievement data from two elementary schools in a school system in Virginia. Achievement Development Scale Scores (ADSS) from the California Achievement Test were collected in grades 2 through 6 at the end of the school years 1977-78 and 1978-79. The tests were administered in May of both years.

A sample of classes from grades 4, 5, and 6 in the two schools was selected. These grades were the common grades for which data were available. Classes were selected using a random sampling design, stratified by grade level and school. The strata were 3 grade levels and 2 schools, yielding 6 cells. Since grade and school populations were approximately equal, three classes were selected for each cell, for a total sample size of 18 classes.

After the sample of 18 classes had been drawn from the three grades in Schools A and B, the data were analyzed using analysis of covariance and within-class regression. The purpose of this part of the study was to explore the utility of the technique of within-class regression using real data and to compare the results to an analysis of covariance design.

Results

The data which were generated closely fit the mathematical model described earlier. In the first condition, correlations between the pretest and posttest were similar and clustered around .75. Standard deviations were also similar, clustering around 2.00. The correlations were varied in the second conditions by varying the weighting of the true score component. The different weightings caused the standard deviations of the posttests to also vary. In order to reset the standard deviations to their original values, the scores were multiplied by the ratios of the old standard deviation over the new standard deviation. This multiplication caused the standard deviations to be reset to the original values. The variances in the third condition were varied to cause a heterogeneity of variance. The data were analyzed under all three conditions using a one-way analysis of covariance design and within-class regression.

The analysis of covariance results under the first condition yielded a statistically significant treatment effect. The F ratio was 5.59 with 19 and 9,979 degrees of freedom,

which was significant at $p < .001$. When the adjusted means were correlated with the pretest means, the correlation coefficient was .921, indicating that adjusted means were still strongly related to pretest levels. Thus, the analysis of covariance yielded biased results.

The results were quite different when the same data were analyzed using the technique of within-class regression. The F ratio was 1.25 with 19 and 9,960 degrees of freedom, which was not significant at $p = .05$. The F ratio correctly indicated a no-treatment effect. Moreover, the residual means were uncorrelated with the pretest means ($r = .00$). Thus, the analysis at the grand mean correctly indicated a no-treatment effect and the residual means were not confounded with the pretest means.

The analysis in the second condition, in which reliability varied across groups, yielded results which were similar to the results in the first condition. The results of the analysis of covariance yielded an F ratio of 5.15 with 19 and 9,979 degrees of freedom, which was significant at the .001 level. The correlation coefficient between the adjusted means and the pretest means was .913, indicating again a confounding between the pretest means and the adjusted means.

The results of the within-class regression analysis were also similar to the results in the first condition. The F ratio was 1.24 with 19 and 9,960 degrees of freedom, which was not significant at the $p = .05$ level. The residual means were also uncorrelated with the pretest means. The results

of the analysis of covariance, when reliability varied across groups, would have led to a false rejection of the null hypothesis while the within-class regression results correctly failed to reject the null hypothesis. Additionally, the stronger treatment effects would have been attributed to the groups with higher pretest mean scores through the use of analysis of covariance.

Variances in the posttest were allowed to vary across groups in the third condition causing a violation of the assumptions of homogeneity of variance of analysis of covariance. Stability coefficients were held constant at .75 as in the first condition. When the data were analyzed using analysis of covariance, the F ratio, with 19 and 9,979 degrees of freedom, was 2.46, which was significant at the .01 level. The correlation between the adjusted mean scores and the pretest scores was .921. As in the first two conditions, the analysis of covariance led to a false rejection of the null hypothesis and the adjusted means were confounded with the pretest means.

The results of the within-class regression analysis at the grand mean yielded an F ratio of 8.50 with 19 and 9,960 degrees of freedom, which was significant at the .001 level. However, the residual means were still uncorrelated with pretest means (.000). Since the residual means were independent of pretest means, the results seemed to indicate an unbiased estimate of a treatment effect specifically at the grand mean of the pretest.

The data were also analyzed at points one standard deviation above the pretest grand mean and one standard deviation below the pretest grand mean under all three conditions. The results were consistent in all three conditions. The residual means were always uncorrelated with the pretest means, indicating that treatment effect estimates were independent of pretest levels. However, the statistical test of significance did not work well when conducted away from the grand mean in any condition. The F ratios were always significant at these points. In the second condition, differing reliabilities affected the slopes of the regression lines and affected the treatment effect estimates away from the grand mean. In the third condition, differing variances had the same effects on regression lines. In all conditions, the regression lines became more unstable away from the group means. This phenomenon probably caused an underestimation of the error term conditions resulting in an overestimation of the treatment effects.

As a final part of the study, reading achievement test data from a county school system were analyzed. Eighteen classes were randomly selected from grades four through six. Since the classes came from different grade levels, preexisting differences were substantial. Moreover, descriptive statistics were computed that indicated heterogeneous variances and regressions. Correlation coefficients between pretests and posttests varied across classes from .208 to .874. Standard

deviations varied from 28.4 to 51.4 in the pretest and from 27.2 to 58.0 in the posttest. These statistics indicated that homogeneity of variance and regression assumptions would be violated by the use of analysis of covariance.

When the data were analyzed using analysis of covariance, the results indicated that there was a statistically significant difference between the classes after adjusting for pre-existing differences on the pretest. The F ratio was 4.73 with 17 and 340 degrees of freedom and was statistically significant at $p = .01$. One would conclude from the analysis of covariance that there is a difference in the adjusted means of the classes. Unlike in the Monte Carlo study, the population parameters were unknown and one cannot be sure whether the results were because of real treatment effects or because of biases due to violations of analysis of covariance assumptions.

The within-class regression analysis yielded different results. The reading achievement data were analyzed using the analysis of variance model at the grand mean of the pretest distribution. Since the results of the analysis of variance significance tests were approximately correct at the grand mean in the simulations but were incorrect at the extreme points, the significance test was used only to analyze data at the grand mean. An F ratio of 2.30 with 17 and 324 degrees of freedom was obtained, which was significant at the .01 level. The F ratio indicated that there was a statistically significant difference in the residual means of the classes

after controlling for the preexisting differences in reading achievement. The conclusion that there was a statistically significant treatment effect is the same conclusion reached through the analysis of covariance. However, the conclusions concerning the relative effectiveness of the teachers was quite different between the two methods.

Residual scores were computed and compared to the adjusted mean scores. The two methods yielded different rank orderings of treatment effects (Table 2). The difference in rank orderings indicates that there is not only a theoretical difference in the results of the two methods of analysis but that this difference also affects the relative effectiveness attributed to teachers. Different teachers sometimes received substantially different effectiveness ratings depending on which method of analysis was used. Teacher 3, for example, was rated only average with a rank of 9 out of 18 teachers when analysis of covariance was used. The class started off low and made a large gain, as can be observed from the pretest and posttest mean scores. When within-class regression analysis was used, the same teacher was rated as the most effective teacher. Teacher 14 was ranked 17 by the analysis of covariance method but 2 by the within-class regression method. These differences in rankings indicate that classes with low beginning levels were penalized by the covariance analysis but not by the within-class regression analysis.

Insert Table 2 about here

The residual means and adjusted means were also correlated with the pretest mean scores. The results were very similar to the results from the Monte Carlo section of the study. The correlation between the adjusted mean scores and the residual mean scores was .777, indicating that the adjusted means were still highly correlated with the pretest mean scores. The correlation between the residual mean scores and the pretest mean scores was .000. These results rather clearly indicate that the within-class regression procedure yielded results that were independent of the pretest scores while the adjusted means were still confounded with the preexisting differences between the groups and yielded biased results.

Comparisons of teacher effects were made at the grand mean and at points one standard deviation above the pretest grand mean and one standard deviation below the pretest grand mean. The within-class regression technique sometimes showed teachers to be effective or ineffective at all levels but sometimes showed teachers to be differentially effective at different preexisting achievement levels. These results seem to be realistic results that normally could occur with teacher effectiveness research. The ability of within-class regression to pick up differing treatment effects at different preexisting levels seems to be a major advantage over analysis of covariance procedures that yield only an overall treatment effect.

Summary

The analysis in this study consisted of an analysis of Monte Carlo data and reading achievement data. The Monte Carlo data were analyzed under three conditions. In the first condition, measurement error was present but constant across groups. The analysis of covariance was shown to yield biased results favoring the initially higher groups. These results were consistent with results reported in the literature (Campbell & Erlebacher, 1970; Campbell & Boruch, 1975; Lord, 1960). The within-class regression analysis indicated that this technique is not biased in favor of groups with higher pretest scores.

Measurement error was varied across groups in the second condition. The analysis of covariance results were also biased in favor of the groups with higher pretest scores. The within-class regression results were still independent of the pretest means but the results depended on where on the pretest distribution the analysis was conducted. When the analysis was conducted at the grand mean, the results indicated no treatment effect. However, when the analysis was conducted at points away from the grand mean of the pretest distribution, the results were confounded with the reliability of the measures and pseudo-treatment effects were evident.

In the third condition, measurement error was constant across groups but there was a heterogeneity of variance between the groups. The analysis of covariance indicated, again,

a significant treatment effect but the F ratio was smaller than in the first two conditions. The within-class regression analysis was shown to have the ability to reflect differing treatment effects at different levels of the pretest distribution. It seemed that the within-class regression technique showed correct results at the grand mean. However, in this case, the analysis of data at points away from the pretest mean led to biased results.

In addition to comparing the results of the within-class regression technique to the analysis of covariance results, a statistical test based on an analysis of variance model was tested. The results of using this model to test the significance of the residual means were approximately correct when the test was conducted at the grand mean of the pretest distribution. In both the first and second conditions, the results of the test of significance indicated a no-treatment effect at the grand mean. However, when the test was applied to the points away from the grand mean, the test incorrectly indicated a treatment effect. At the grand mean in the third condition, the significant result appeared to be the result of a real treatment effect at the grand mean of the pretest distribution:

Real data were also analyzed to see if any differences between the two results would be apparent. Reading achievement data from an elementary school age population were analyzed using both methods. The analysis of covariance results were confounded with preexisting conditions as in the simulated

data. The within-class regression results were not confounded with pretest differences and the results were quite different than the analysis of covariance results. Depending on the method of analysis used, different teachers were rated as the most effective. The results of this part of the study indicated that there were not only theoretical differences in the results of the two methods but that these differences affect the effectiveness ratings attributed to teachers of non-equivalent classes.

TABLE 1

J_1		$J_2 \dots \dots \dots J_{20}$			
<u>X</u>	<u>Y</u>	<u>X</u>	<u>Y</u>	<u>X</u>	<u>Y</u>
$I_{1,1,1}$	$I_{1,2,1}$	$I_{2,1,1}$	$I_{2,2,1} \dots$	$I_{20,1,1}$	$I_{20,2,1}$
$I_{1,1,2}$	$I_{1,2,2}$	$I_{2,1,2}$	$I_{2,2,2} \dots$	$I_{20,1,2}$	$I_{20,2,2}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
$I_{1,1,500}$	$I_{1,2,500}$	$I_{2,1,500}$	$I_{2,2,500} \dots$	$I_{20,1,500}$	$I_{20,2,500}$

I = subject

J = group

X = pretest score

Y = posttest score

TABLE 2
Comparison of Pretest, Posttest, Adjusted,
and Residual Means
(Reading Achievement Data)

Class	Pretest Mean	Posttest Mean	Adjusted Mean	Residual Mean
1	300.12 (17)	324.06 (18)	372.55 (18)	-32.369 (18)
2	406.43 (5)	439.73 (6)	416.43 (6)	- 5.501 (13)
3	264.40 (18)	340.40 (16)	413.02 (9)	19.229 (1)
4	341.38 (15)	371.25 (14)	391.88 (13)	- 8.777 (16)
5	367.83 (10)	404.00 (9)	406.76 (10)	7.715 (3)
6	366.67 (11)	399.81 (10)	403.36 (11)	4.864 (8)
7	396.64 (7)	440.29 (5)	423.59 (4)	6.202 (5)
8	354.80 (12)	378.53 (13)	390.10 (15)	- 3.144 (12)
9	435.00 (2)	476.08 (2)	433.48 (2)	5.689 (6)
10	449.32 (1)	502.32 (1)	450.05 (1)	4.921 (7)
11	412.89 (4)	447.11 (4)	419.44 (5)	- 6.795 (14)
12	368.55 (9)	389.36 (11)	391.64 (14)	- 9.589 (17)
13	405.56 (6)	438.47 (7)	414.73 (7)	- 6.932 (15)
14	303.48 (16)	331.56 (17)	377.78 (17)	12.671 (2)
15	393.73 (8)	429.81 (8)	415.08 (8)	3.157 (9)
16	434.27 (3)	470.23 (3)	428.12 (3)	2.157 (10)
17	351.00 (13)	388.39 (12)	402.52 (12)	7.221 (4)
18	342.50 (14)	369.94 (15)	389.82 (16)	- 1.266 (11)

(Ranks in Parenthesis)

References

- Burstein, L., Linn, R. L., & Capell, F. J. Analyzing multi-level data in the presence of heterogeneous within-class regressions. Journal of Educational Statistics, Winter, 1978, 3, 347-383.
- Campbell, D. T., & Boruch, R. F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), Evaluation and experiment: Some critical issues in assessing social programs. Seattle, Washington: Academic Press, Inc., 1975.
- Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), The disadvantaged child (Vol. 3). New York: Breimer/Mazel, 1971.
- Cronbach, L. J., Rogosa, D. R., Floden, R. E., & Price, G., Analysis of covariance: Angel of salvation or temptress and deluder? Stanford, Ca.: Stanford University, 1976 (ERIC Document Reproduction Service No. ED 128 361).
- Elashoff, J. D. Analysis of covariance: A delicate instrument. American Educational Research Journal, 1969, 6, 383-402.

- Furby, L. Interpreting regression toward the mean in developmental research. Developmental Psychology, 1973, 8, 172-9.
- Lord, F. M. Large scale covariance analysis when the control variable is fallible. Journal of the American Statistical Association, 1960, 55, 307-21.
- Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change. Madison: University of Wisconsin Press, 1963.
- Lord, F. M. Statistical adjustments when comparing pre-existing groups. Psychological Bulletin, 1969, 72, 336-337.
- Porter, A. C. How errors of measurement affect anova, regression analysis, ancova and factor analysis, Ypsilana, Mich.: Michigan State University, 1971 (ERIC Document Reproduction Services No. ED 050 172).
- Pothoff, R. F. On the Johnson-Neyman technique and some extensions thereof. Psychometrika, 1964, 29, 241-256.